WHAT IS CLAIMED IS:

Sub A7

1. A method for distributing a streaming multimedia (SM) object in a network having a content server which hosts SM objects for distribution over said network through a plurality of helpful servers (HSs) to a plurality of clients, said method comprising:

calculating at said content server a server hotness rating for said SM objects hosted thereon;

performing a categorization process, wherein each of said SM objects hosted by said content server are categorized into one of a plurality of server hotness categories based on each of said SM object's calculated server hotness rating; and

multicasting from said content server at least one of said SM objects hosted thereon to a fraction of said plurality of HSs in the network, said fraction being determined according to said SM object's hotness category.

2. A method as recited in claim 1, further comprising the step of associating a fraction to each of said plurality of predetermined hotness categories before multicasting said SM objects.

3. A method as recited in claim 1, wherein the server hotness rating for each of said SM object's hosted by said content server is calculated as the sum of a plurality of helper hotness ratings, wherein the helper hotness rating for an SM object hosted by one of said plurality of HSs is defined as a total number of client requests for said SM object requested from said one of said plurality of HSs divided by a time span in which said client requests are received.

Sub A7

4. A method as recited in claim 1, wherein each of said plurality of server hotness categories are defined by a lower server hotness rating value and an upper server hotness rating value.

5. A method for distributing at least a portion of a streaming multimedia (SM) object in a network having a content server which hosts SM objects for distribution over said network through a plurality of HSs to a plurality of clients, said method comprising:

calculating at said content server a server hotness rating for said SM objects hosted thereon;

performing a categorization process, wherein each of said SM objects hosted by said content server are categorized into one of a plurality of server hotness categories based on each of said SM object's calculated server hotness rating; and

multicasting from said content server a fraction of said SM object to said plurality of HSs in the network, said fraction being determined according to said SM object's hotness category.

6. A method as recited in claim 5, wherein the server hotness rating for each of said SM object's hosted by said content server is calculated as the sum of a plurality of helper hotness ratings, wherein the helper hotness rating for an SM object hosted by one of said plurality of HSs is defined as a total number of client requests for said SM object requested from said one of said plurality of HSs divided by a time span in which said client requests are received.

Sub A7

7. A method as recited in claim 5, further comprising the step of associating a fraction to each of said plurality of predetermined hotness categories before multicasting said SM objects.

8. The method of claim 5, wherein each of said plurality of server hotness categories are defined by a lower server hotness rating value and an upper server hotness rating value.

9. A method for distributing at least a portion of a streaming media (SM) object in a network having a content server which hosts SM objects for distribution over said network through a plurality of HSs to a plurality of clients, said method comprising:

calculating at said content server a server hotness rating for said SM objects hosted thereon;

performing a categorization process, wherein each of said multiple SM objects being hosted by said content server are categorized into one of a plurality of server hotness categories based on each of said SM object's calculated server hotness rating; and

requesting by one of said plurality of HSs in the network a fraction said SM object from said content server, said fraction being determined according to said SM object's hotness category.

10. A method as recited in claim 1, further comprising the step of associating a fraction to each of said plurality of predetermined hotness categories before requesting said SM objects.

11. A method as recited in claim 9, wherein the server hotness rating for each of said SM object's hosted by said content server is calculated as the sum of a plurality of helper hotness ratings as the sum of a plurality of helper hotness ratings, wherein the helper hotness rating for an SM object hosted by one of said plurality of HSs is defined as a total number of client requests for said SM object made to said one of said plurality of HSs divided by a time span in which said client requests are received.

11. A method as recited in claim 9, wherein each of said plurality of predetermined server hotness categories are defined by a lower server hotness rating value and an upper server hotness rating value.

13. A method for distributing a streaming multimedia (SM) object in a network having a content server which hosts SM objects for distribution over said network through a plurality of HSs to a plurality of clients, said method comprising:

calculating at said content server a server hotness rating for said SM objects hosted thereon;

performing a categorization process, wherein each of said SM objects hosted by said content server are categorized into one of a plurality of server hotness categories based on said SM object's server hotness rating; and

Sub A' 7

multicasting from said content server a first fraction of said SM object to a second fraction of said plurality of HSs in the network, said first and second fraction being determined according to said SM object's hotness category.

14. A method for storing a streaming multimedia (SM) object in a network having a content server which hosts SM objects for distribution over said network through a plurality of helper servers (HS) to a plurality of clients, said streaming media object being comprised of a plurality of successive time-ordered chunks, wherein a chunk is further comprised of an integer multiple number of HS storage disk blocks,

said method comprising:

i) receiving a streaming media object;

ii) determining whether sufficient disk space is available on said at least one of said plurality of HSs to store said received SM object;

iii) storing said received SM object at said at least one of said plurality of HSs if it is determined that there is sufficient disk space; and

iv) performing the following steps, if it is determined that there is insufficient disk space available to store the received SM object:

a) identifying at least one SM object from among a plurality of SM objects hosted by said HS which is not in use and has an access time which is least recent, wherein said access time corresponds to a time when said SM object was last requested; and

b) replacing chunks of said identified at least one SM object

Sub A7

corresponding to said received SM object to store at least a portion of said received SM object.

15. A method as recited in claim 14, wherein the step of replacing chunks further comprises the step of replacing the chunks having the largest associated time-stamp value of said at least one identified SM object.

16. A method as recited in claim 15, wherein the step of replacing said chunks having the largest time-stamp value further comprises the step of replacing said chunks having the largest time-stamp value on a disk block by disk block basis starting from the end of the chunk.

17. A method for storing a streaming media (SM) object in a network having a content server which hosts SM objects for distribution over said network through a plurality of helper servers (HSs) to a plurality of clients, said SM object being comprised of a plurality of successive time-ordered chunks, wherein a chunk is further comprised of some integer multiple number of disk blocks, said method comprising:

i) receiving said SM object;

ii) determining whether there is disk space available on said one of said plurality of HSs;

iii) storing said SM object at said at least one HS if it is determined that there is sufficient disk space available; and

Sub A'7

iv) performing the following steps, if it is determined that there is insufficient disk space available:

a) composing a set of SM objects from among a plurality of SM objects stored on said disk space whose access time is determined to be least recent, where said access time corresponds to a time when said SM object was last requested; and

b) replacing chunks of said SM objects belonging to said composed set with chunks of said received SM object.

18 A method as recited in claim 17, wherein the replacing step further comprises the step of replacing a chunk having an associated highest time-stamp value from each of said SM objects belonging to said composed set in a round-robin basis.

19. A method as recited in claim 17, wherein said composed set is formed by including only SM objects having a helper hotness rating below a predefined threshold.

20. A method for storing a streaming media (SM) object in a network having a content server which hosts SM objects for distribution over said network through a plurality of helper servers (HSs) to a plurality of clients, said SM object being comprised of a plurality of successive time-ordered chunks, wherein a chunk is further comprised of some integer multiple number of disk blocks, said method comprising:

i) receiving said SM object at one of said plurality of HSs in said

Sub A'7

network; and

ii) randomly determining a fraction and storing said fraction of said SM object at said one of said HSs in the network.

21. A method as recited in claim 20, wherein the step of randomly determining said fraction said fraction comprises the steps of:

i) estimating at said one of said plurality of HSs, the number of HSs in said network, K;

ii) calculating a value, P, that is inversely proportional to the value K;

iii) generating a random value, R, for a received chunk of said received SM object;

iv) comparing said generated random value, R, with said calculated value P;

v) caching the received chunk, if it is determined that the generated random value, R, is less than the calculated value P; and

vi) discarding the received chunk, if it is determined that the generated random value, R, is greater than the calculated value P.

22. The method according to claim 21, wherein the step of generating the random value R is performed by utilizing the IP address of said one of said plurality of HSs as a seed.

Sub A7

23. A method for storing a streaming media (SM) object in a network having a content server which hosts SM objects for distribution over said network through a plurality of helper servers (HSs) to a plurality of clients, said SM object being comprised of a plurality of successive time-ordered chunks, wherein a chunk is further comprised of some integer multiple number of disk blocks, said method comprising:

i) requesting said SM object at one of said plurality of HSs in said network;

ii) receiving said SM object at one of said plurality of HSs in said network; and

iii) randomly determining a fraction and storing said fraction of said SM object at said one of said HSs in the network.

24. A method as recited in claim 23, wherein the step of randomly determining said fraction said fraction comprises the steps of:

i) estimating at said one of said plurality of HSs, the number of HSs in said network, K;

ii) calculating a value, P, that is inversely proportional to the value K;

iii) generating a random value, r, for a received chunk of said received SM object;

iv) comparing said generated random value, R, with said calculated value P;

v) caching the received chunk, if it is determined that the generated

random value, R, is less than the calculated value P; and

vi) discarding the received chunk, if it is determined that the generated random value, R, is greater than the calculated value P.

25. The method according to claim 24, wherein the step of generating the random value R is performed by utilizing an arrival time as the seed of said SM object at said one of said plurality of HSs.

26. A method for storing a streaming media (SM) object in a network having a content server which hosts SM objects for distribution over said network through a plurality of HSs to a plurality of clients, said SM object being comprised of a plurality of successive time-ordered chunks, wherein a chunk is further comprised of some integer multiple number of disk blocks, said method comprising:

i) receiving said SM object at one of said plurality of HSs;

ii) determining whether there is disk space available on said one of said plurality of HSs to store said SM object;

iii) storing said SM object at said one of said plurality of HSs if it is determined that there is sufficient disk space available; and

iv) performing the following steps, if it is determined at the determining step that there is insufficient disk space available:

a) composing a set of streaming media objects from among a plurality of streaming media objects stored on said disk space, whose access time

Sub A7

is determined to be above a predetermined threshold, wherein said access time is a time when said SM object was last retrieved;

b) assigning a unique integer value to each chunk of the composed set of SM objects based at least one of a name of the SM object, a time-stamp of the chunk, and a last access time of the chunk;

c) generating a random value, R, having a range between one and the largest unique integer value assigned; and

d) replacing a stored chunk having said assigned integer value which corresponds to the generated random value with a chunk of said received SM object.

27. A method for servicing a client request for a streaming media object in a network having a content server which hosts SM objects for distribution over said network through a plurality of HSs to a plurality of clients, said SM object being comprised of a plurality of successive time-ordered chunks, said method comprising:

i) receiving a request for at least a portion of an SM object at a local HS, said request including a playback starting time;

ii) determining whether a chunk having a starting time equal to the requested playback starting time resides on a disk associated with said local HS;

iii) delivering to one of said plurality of clients said one or more chunks having an associated starting time equal to the requested playback starting time from said local HS and updating said playback starting time ;

Sub A7

iv) identifying a server in said network storing at least M chunks which have a starting time equal to said updated playback starting time and which satisfies a minimum cost criterion, if it is determined that said updated playback starting time is not equal to an end-time of said SM object;

v.) receiving said at least M chunks from said identified server at said at local HS until it is determined that a next chunk to be retrieved is stored at said local HS;

28. A method for determining an optimal sequence of cache accesses to service a request for a streaming media (SM) object in a network including a content server which hosts SM objects for distribution over said network through a plurality of helper servers (HSs) to a plurality of clients, the SM objects being comprised of a plurality of successive time-ordered chunks, the method comprising the steps of:

(i) receiving a request at a local HS for an SM object, said request including a requested playback starting time;

(ii) determining whether a chunk of an SM object having a starting time equal to the requested playback starting time is stored at said local HS;

(iii) retrieving said one or more chunks of said SM object when step (ii) is satisfied and updating the requested playback starting time;

(iv) identifying a minimum cost server in the network which stores at least M chunks of said SM object having a starting time equal to the updated playback starting time;

(v) retrieving said at least M chunks on a chunk by chunk basis from said identified minimum cost server and said local HS;

(vi) retrieving said next chunk from said local HS and updating the requested playback starting time;

(vii) repeating step (ii) through step (vi) until the entire streaming media object has been retrieved.

29. The method of claim 28, wherein at step (v), a chunk is always retrieved from said local HS if it is available.

30. The method of claim 29 further comprising the step of returning chunks of said requested SM object to one of said plurality of clients as they are retrieved at said local HS.